

IOWG Session

Feb 7th

Tuesday Afternoon

12:45-1:25 Rick Stevens, TeraGrid

1:25-2:15 BRC update/discussion

Break

2:30-2:45 Ross Overbeek, Related Genomes

2:45-4:00 SOP discussion

The Search for Ground Truth

- TIGR + hegemony → 234

The Search for Ground Truth

- TIGR + hegemony → 234
- TIGR + hubris → 316

The Search for Ground Truth

- TIGR + hegemony → 234
- TIGR + hubris → 316
- Owen + White + incompetent → 108,000

The Search for Ground Truth

- TIGR + hegemony → 234
- TIGR + hubris → 316
- Owen + White + incompetent → 108,000
- Bruno + Sobral + incompetent → 28



BRC-central Usage

BRC Central				
Month	Unique Visitors	Number of visits	Page visits	Bandwidth
Nov 2005	78	98	240	8.75 MB
Dec 2005	105	147	316	13.51 MB
Jan 2006	53	78	261	14.39 MB

BRC Central - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.brc-central.org/cgi-bin/prok_manatee/brc-central/brc_central Go



BRC Central
Bioinformatics Resource Center

Home
Search Tools
FTP
Organisms
Software
Links
Contact Us
About Us

BRC Central

[ApiDB](#) [BioHealthBase](#) [ERIC](#) [NMPDR](#) [Pathema](#) [Patric](#) [VBRC](#) [VectorBase](#)

► **BRC Central** A repository linking to eight Biodefense Resource Centers (BRCs) sponsored by the NIAID. The BRCs are providing web-based resources to scientific community conducting basic and applied research on organisms considered potential agents of biowarfare or bioterrorism or causing emerging or re-emerging diseases.

These centers support existing and newly developed techniques for bioinformatic analysis aimed at obtaining a deeper understanding of the fundamental biology of a specific set of pathogenic organisms, and efforts to counter the threats posed by these pathogens.

► **ApiDB -- Apicomplexan database** portal to several sites including *Toxoplasma gondii*, *Cryptosporidium parvum*.

► **BioHealthBase -- The Biodefense/Public Health DataBase** focuses on data about six priority pathogens to help fill in gaps in genomic and other data critical to scientific researchers. The six pathogens are: *Giardia lamblia* parasite, *Mycobacterium tuberculosis*, Influenza virus, *Francisella Tularensis*, *Microsporidia* parasites and *Ricinus communis* (castor bean).

► **ERIC -- Enteropathogen Resource Integration Center** resource for five members of the family Enterobacteriaceae including: Diarrheagenic *E. coli*, *Shigella*, *Salmonella*, *Yersinia enterocolitica*, *Yersinia pestis*.

► **NMPDR -- National Microbial Pathogen Data Resource Center** the physiology of the pathogens, to clarify the detailed variations that determine phenotype, and to develop consistent interpretations of functional data with special focus on *Staphylococcus aureus*, pathogenic vibrios, *Listeria monocytogenes*, *Campylobacter jejuni*, *Streptococcus pyogenes*, and *Streptococcus pneumoniae*.

► **Pathema -- functional assignments, metabolic reconstructions and transporters** for: *Bacillus anthracis*, *Clostridium botulinum*, *Burkholderia mallei*, *Burkholderia pseudomallei*, *Clostridium perfringens* and *Entamoeba histolytica*.

► **PATRIC -- PathoSystems Resource Integration Center** providing a comprehensive and accurate web-based resource for genomic and associated information on a number of important human pathogens including *Brucella*, *Coxiella burnetii*, *Rickettsia*, and the following viruses: Caliciviruses, Coronaviruses, Hepatitis A, Hepatitis E and Rabies.

► **VBRC -- Viral Bioinformatics Resource Center** encompassing the viral families: Arenaviridae, Bunyaviridae, Flaviviridae, Filoviridae, Paramyxoviridae, Poxviridae, and Togaviridae.

► **VectorBase -- Invertebrate Vectors of Human Pathogens** including *Anopheles gambiae*, *Anopheles aegypti* and other organisms

BRC Central - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.brc-central.org/cgi-bin/prok_manatee/brc-central/brc_central Go



BRC Central

Home

Search Tools

FTP

Organisms

Software

Links

Contact Us

About Us

ApiDB BioHealthBase ERIC NMPDR Pathema Patric **VBRC** VectorBase

Keyword Search

Attribute Search

Organism Search

Taxonomy Browser

hyper linking to eight Biodefense Resource Centers (BRCs) sponsored by the NIAID. The web-based resources to scientific community conducting basic and applied research on potential agents of biowarfare or bioterrorism or causing emerging or re-emerging

These centers support existing and newly developed techniques for bioinformatic analysis aimed at obtaining a deeper understanding of the fundamental biology of a specific set of pathogenic organisms, and efforts to counter the threats posed by these pathogens.

► **ApiDB -- Apicomplexan database** portal to several sites including *Toxoplasma gondii*, *Cryptosporidium parvum*.

► **BioHealthBase -- The Biodefense/Public Health DataBase** focuses on data about six priority pathogens to help fill in gaps in genomic and other data critical to scientific researchers. The six pathogens are: *Giardia lamblia* parasite, *Mycobacterium tuberculosis*, Influenza virus, *Francisella Tularensis*, *Microsporidia* parasites and *Ricinus communis* (castor bean).

► **ERIC -- Enteropathogen Resource Integration Center** resource for five members of the family Enterobacteriaceae including: Diarrheagenic *E. coli*, *Shigella*, *Salmonella*, *Yersinia enterocolitica*, *Yersinia pestis*.

► **NMPDR -- National Microbial Pathogen Data Resource Center** the physiology of the pathogens, to clarify the detailed variations that determine phenotype, and to develop consistent interpretations of functional data with special focus on *Staphylococcus aureus*, pathogenic vibrios, *Listeria monocytogenes*, *Campylobacter jejuni*, *Streptococcus pyogenes*, and *Streptococcus pneumoniae*.

► **Pathema -- functional assignments, metabolic reconstructions and transporters** for: *Bacillus anthracis*, *Clostridium botulinum*, *Burkholderia mallei*, *Burkholderia pseudomallei*, *Clostridium perfringens* and *Entamoeba histolytica*.

► **PATRIC -- PathoSystems Resource Integration Center** providing a comprehensive and accurate web-based resource for genomic and associated information on a number of important human pathogens including *Brucella*, *Coxiella burnetii*, *Rickettsia*, and the following viruses: Caliciviruses, Coronaviruses, Hepatitis A, Hepatitis E and Rabies.

► **VBRC -- Viral Bioinformatics Resource Center** encompassing the viral families: Arenaviridae, Bunyaviridae, Flaviviridae, Filoviridae, Paramyxoviridae, Poxviridae, and Togaviridae.

► **VectorBase -- Invertebrate Vectors of Human Pathogens** including *Anopheles gambiae*, *Anopheles aegypti* and other organisms



BRC Organisms

[Home](#) | [FTP](#) | [Links](#) | [Contact Us](#)

Category	ApiDB	BioHealthBase	ERIC	NMPDR
A	-	<i>Francisella tularensis</i>	<i>Yersinia pestis</i>	-
B	<i>Toxoplasma gondii</i> , <i>Cryptosporidium parvum</i> , <i>Plasmodium phylum</i>	<i>Giardia lamblia</i> , <i>Microsporidia</i> , <i>Ricinus communis</i>	<i>Diarrheagenic E. coli</i> , <i>Yersinia enterocolitica</i> , <i>Shigella</i> , <i>Salmonella</i>	<i>Staphylococcus aureus</i> , <i>Pathogenic Vibrios</i> , <i>Listeria monocytogenes</i> , <i>Campylobacter jejuni</i> , <i>Streptococcus pyogenes</i> , <i>Streptococcus pneumoniae</i>
C	-	<i>Mycobacterium tuberculosis</i> , <i>Influenza Virus</i>	-	-

Category	Pathema	Patric	VectorBase	VBRC
A	<i>Bacillus anthracis</i> , <i>Clostridium botulinum</i>	-	-	<i>Variola major Virus</i> , <i>Arenavirus</i> , <i>Hanta Virus</i> , <i>Rift Valley Fever Virus</i> , <i>Ebola Virus</i> , <i>Marburg Virus</i> , <i>Dengue Virus</i>
B	<i>Burkholderia mallei</i> , <i>Burkholderia pseudomallei</i> , <i>Clostridium perfringens</i> , <i>Entamoeba histolytica</i>	<i>Rickettsiae</i> , <i>Brucella</i> , <i>Coxiella burnetii</i> , <i>Calicivirus</i> , <i>Hepatitis A Virus</i>	-	<i>California encephalitis group Virus</i> , <i>Kyansanar forest disease Virus</i> , <i>Omsk hemorrhagic fever Virus</i> , <i>West Nile Virus</i> , <i>Alphavirus</i>
C	-	<i>Rabies Virus</i> , <i>Coronavirus</i>	<i>Anopheles gambiae</i> , <i>Aedes aegypti</i> , <i>Anopheles gambiae</i> , <i>Culex pipiens</i> , <i>Ixodes scapularis</i>	<i>Hantaan Virus</i> , <i>Puumala Virus</i> , <i>Crimean-Congo hemorrhagic fever Virus</i> , <i>Yellow fever Virus</i> , <i>Tick-borne Encephalitis</i> , <i>Nipah Virus</i> , <i>Equine morbillivirus</i>

Index of ftp://ftp.brc-central.org/ - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

ftp://ftp.brc-central.org/

Go

Index of ftp://ftp.brc-central.org/

[Up to higher level directory](#)

ApiDB

10/25/2005 3:35:00 PM

BHB

10/18/2005 3:16:00 PM

ERIC

8/30/2005 3:25:00 PM

NMPDR

8/16/2005 4:36:00 PM

PATRIC

1/26/2005 2:39:00 PM

TIGR

2/2/2006 5:48:00 PM

VBRC

1/18/2006 3:52:00 PM

VectorBase

2/2/2006 2:23:00 PM

Index of ftp://ftp.brc-central.org/PATRIC - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

ftp://ftp.brc-central.org/PATRIC/

Go

Index of ftp://ftp.brc-central.org/PATRIC

[Up to higher level directory](#)

Brucella species

1/26/2005 2:38:00 PM

Brucella species.tar.gz

6928 KB 1/26/2005 2:39:00 PM

Caliciviruses

1/26/2005 2:43:00 PM

Caliciviruses.tar.gz

205 KB 1/26/2005 2:39:00 PM

Coronaviruses

1/26/2005 3:01:00 PM

Coronaviruses.tar.gz

2218 KB 1/26/2005 2:39:00 PM

Coxiella burnetii

1/26/2005 3:10:00 PM

Coxiella burnetii.tar.gz

1141 KB 1/26/2005 2:39:00 PM

HepatitisA viruses

1/26/2005 3:10:00 PM

HepatitisA viruses.tar.gz

33 KB 1/26/2005 2:39:00 PM

HepatitisE viruses

1/26/2005 3:11:00 PM

HepatitisE viruses.tar.gz

128 KB 1/26/2005 2:39:00 PM

Lyssavirus

1/26/2005 3:11:00 PM

Lyssavirus.tar.gz

57 KB 1/26/2005 2:39:00 PM

README.txt

11 KB 1/17/2006 9:58:00 PM

Rickettsia species

1/26/2005 3:12:00 PM

Rickettsia species.tar.gz

4579 KB 1/26/2005 2:39:00 PM



BRC Central - Software - Mozilla Firefox


File Edit View Go Bookmarks Tools Help

← → ↺ × 🏠

http://www.brc-central.org/tdb/prok_manatee/brc-central/software.shtrr

Go





Home

Search Tools

FTP

Organisms

Software

Links

Contact Us

About Us

BRC Central Software

MANATEE

Manual Annotation Tool

Manatee is a web-based gene evaluation and genome annotation tool that can view, modify, and store annotation for prokaryotic and eukaryotic genomes. The Manatee interface allows biologists to quickly identify genes and make high quality functional assignments using a multitude of genome analyses tools. These tools consist of, but are not limited to GO classifications, BER and blast search data, paralogous families, and annotation suggestions generated from automated analysis. Manatee is funded by The National Institute of Allergy and Infectious Diseases (NIAID) as part of the NIAID Bioinformatics Resource Centers (BRC) for Biodefense and Emerging or Re-Emerging Infectious Diseases.

TIGRDetails

Sybil

Web-based Software for Comparative Genomics

Sybil is a web-based software package for comparative genomics. It was developed by the Bioinformatics department at The Institute for Genomic Research (TIGR) and is funded by The National Institute of Allergy and Infectious Diseases (NIAID) as part of the NIAID Bioinformatics Resource Centers (BRC) for Biodefense and Emerging or Re-Emerging Infectious Diseases. The primary goal of the Sybil software package is to supply online comparative analysis tools for the Pathema web site (a Bioinformatics Resource Center located at TIGR), which will provide in-depth curatorial and comparative analysis of several pathogens.

TIGRDetails

WorkFlow

Annotation Pipeline Monitor

The Institute for Genomic Research (TIGR) has many process pipelines that need to be created, executed, and monitored on an on-going basis. Each pipeline may include multiple discrete process that can be executed either sequentially or in parallel. To reduce manual intervention, and streamline the process flow, TIGR's Annotation software team has designed a system called Workflow that can be used to build, run, and monitor such process pipelines or workflows. WorkFlow is funded by The National Institute of Allergy and Infectious Diseases (NIAID) as part of the NIAID Bioinformatics Resource Centers (BRC) for Biodefense and Emerging or Re-Emerging Infectious Diseases.

TIGRDetails

Software page

- [example](#)

Software Tags

- Software Name
- Institution
- Description
- Headline
- Creator
- Contact
- Web Site
- Open Source
- Download Site
- Current Version
- Current Version Release Date
- Relevant Publications
- Platforms
- Requirements
- Other Information
- Keywords
- Listing Last Updated

Software Tags

- Software Name
- Institution
- Description
- Headline
- Creator
- Contact
- Web Site
- Open Source
- Download Site
- Current Version
- Current Version Release Date
- Relevant Publications
- Platforms
- Requirements
- Other Information
- Keywords
- Listing Last Updated
- Proposal: adopt software submission xml
- Format
- Tags
- Submission strategy

Proposal: feature versioning

- GFF3 contains a `stable_id` attribute
- Format: `accession.version` (e.g., like Genbank)
- Applies to gene (and other genomic) features
- Currently, the `stable_id` attribute is optional for all features
- proposal: mandatory `stable_id` for all gene features
- proposal: accessioning/numbering is responsibility of each BRC
- proposal: changes to underlying sequence spawn version changes

Note: A different mechanism deals with other types of changes

Proposal: data set versioning

- data sets version is created upon BRC Central deposition
- version number specified in GFF3 file
- associated with each genome
- multiple submission (eg. file/chromosome), get same version number
- allows for:
 - independent updates to the GFF3 data
 - capability to retrieve old data sets
- proposal: versioning is responsibility of each BRC
- proposal: storage of old data set is responsibility of each BRC
- proposal: automated check at BRC-central
 - ensure the version# is valid
 - doesn't conflict with a previous submission.

Proposal: tracking feature changes

- Change examples:
 - concept: "Last modified"
 - not: "Show full edit history"
 - changes to annotation
 - new annotations
 - features additions
- This is not versioning. Not all changes result in new accession version.
- For each new data set submission:
 - should come with a change log file to give best effort in detailing change
 - data stamps
 - elaboration of the "Whats New" RSS feed
- Proposal: **any** change triggers new version number for **data set**.

Expected capability for users

- “Retrieve annotations that have been modified...”
- “When was the last time Anthrx strain 12 was last submitted?”
- “I made primers for a gene 2 years ago, where is it?”

brc-central: new features

New features: Round Trip Validation

<u>Center</u>	<u>Ontology_term</u>	<u>ec_number</u>	<u>description (gene_name)</u>	<u>Name (locus)</u>	<u>gene_symbol</u>	<u>organism_name</u>	<u>Dbxref:taxon</u>
ApiDB	0	0	12831	13411	0	1454	1454
BHB	0	0	23003	30357	0	58	58
ERIC	0	2402	102106	104864	40410	947	947
NMPDR	0	0	63878	0	0	292	292
PATRIC	0	0	23991	0	5447	354	354
TIGR	67085	28469	149139	149139	45363	1309	1309
VBRC	0	0	5119	5119	0	414	414
VectorBase	0	0	0	0	0	6	6



GFF3

- Other issues
 - Ev codes → later
 - Other?

Brc-central new features

- Schedule of potential meetings
- Meetings BRCs are attending
- Round trip validation
- Blast searches
- Relationalize
- API

Decoration w/ datatypes

- pathInfo
- BHB
- OrthoMCL
- Sub-systems
- Epitopes (Immune Epitope DB and Analysis Resource)

Improve?



Suggestions for new features

- Searching
- Improved usage statistics
- Notifications
- History management

Break for Ross

Standard Operational Procedures

Feb 7th

Management by Standard Operational Procedures (SOPs)

Software, databases and interfaces
perform routine processes.

Goal: create reproducible and consistent
annotation.

Management by Standard Operational Procedures

Scripts/Pipelines/Automated Analysis:

- encourages rigorous application of processes.
- while curation of data is possible.
- it is still not enough.

The *reasoning* for why processes are applied is not explicit in the system.

The specific points of evaluation need to be identified, and documented.

Benefits of SOPs

- Constructive competition
- Human resources/MGT
 - Chaos is demoralizing
 - Enforces clarification of thought
 - Pumpkinmorphogenesis
 - Essential for training
- Scale-up
 - Knowledge of required resources
 - Coordination
 - Identifies specialized working groups

Benefits of SOPs

- Publication
 - Scientifically defensible
 - Establishes standards in the community
- Retooling the flow diagram
- Meaningful tracking of output
- Essential element of project management:
 - Exit criteria
 - Who evaluates the exit criteria



SOP examples

Engine assembly

El tigre SOPs

SOPs are *not*...

- A process diagram
- A software program or pipeline
- A user manuals
- A scientific publication (e.g., “we did our assignments according to PNAS:342...”)
- Controlled vocabularies
 - Note: `ev_types`

SOP sharing

- Sharing SOPs will help
- Caveats, output still dependent on:
 - Hangovers
 - People
 - Training/increases in expertise
 - Quality of input information
 - Ergonomics

SOP Sharing

- Converge on *types*
 - Nucleotide level
 - Protein level
 - Automated assignment
 - Frameshifts
 - Vaccine targets



Proposal: Decoration w/ datatypes

- pathInfo
- BHB
- OrthoMCL
- Sub-systems
- Epitopes (Immune Epitope DB and Analysis Resource)
- SOPS: Nucleotide-level, protein-level, frameshifts
 - BRC
 - Date stamp
 - SOP version
 - NOTE: ev_codes \leftrightarrow SOPS

Proposal: Decoration w/ datatypes

- Protein-level, implications
 - I'm talking GO
 - \$\$
 - Where to get help
 - ev_codes $\leftarrow \sim \rightarrow$ SOPs